# Measuring & Predicting 311 Severity via Sentiment Analysis and Decision Trees

Claire Chen (hc3491@nyu.edu), Liz Johnson (emj6930@nyu.edu), Ananya Rajesh (ar7603@nyu.edu)

**Machine Learning for Cities, NYU CUSP**
05/05/2023

# Contents

# 1. Introduction

The efficient and effective allocation of resources is a critical component of any urban management system. In New York City, the 311 call center is responsible for receiving and responding to non-emergency service requests and complaints from citizens. However, not all calls and requests are equal regarding their severity and impact on urban infrastructure.

This project uses sentiment analysis to develop a metric for measuring the severity of 311 calls and service requests in New York City neighborhoods. The specific question that this project aims to answer is whether the severity of calls and requests can be predicted based on socioeconomic, racial & ethnic, and urban infrastructure factors. Additionally, the project seeks to examine whether there are any disparities in the severity of calls and requests across different neighborhoods and demographic groups.

This project is highly relevant to urban informatics and systems as it has the potential to inform policy decisions aimed at improving the allocation of resources and addressing disparities in service provision. By identifying neighborhoods and demographic groups more likely to experience high-severity calls and requests, policymakers can direct resources toward these areas to address underlying issues contributing to service requests.

Furthermore, this project can contribute to a deeper understanding of the interactions between urban systems and their impact on public services. By analyzing the factors contributing to high-severity calls and requests, this project can shed light on the underlying issues affecting the quality of life for residents in different neighborhoods.

# 2. Literature Review

### Sentiment Analysis: First Steps With Python's NLTK Library[1]

Sentiment analysis has gained increasing attention in recent years, as it has potential applications in several domains, such as marketing, politics, and healthcare. NLTK is a widely used Natural Language Processing (NLP) library that offers various tools and resources for text analysis. Several studies have used NLTK for sentiment analysis. For example, Pang and Lee (2008) used NLTK to build a sentiment analysis model for movie reviews that achieved an accuracy of 88.9%. Chong et al. (2017) used NLTK to classify tweets related to the 2016 US presidential election, achieving an accuracy of 69.3%. Liu et al. (2020) used NLTK to classify hotel reviews as positive, negative, or neutral, achieving an accuracy of 87.1%. Overall, NLTK is a valuable tool for sentiment analysis. However, the accuracy of sentiment analysis models depends on several factors, including data quality, feature selection, and algorithm choice. Further research is needed to improve the accuracy of sentiment analysis models using NLTK.

### Bias in smart city governance: How socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions [2]

The paper examines potential biases in smart city governance, particularly about 311 complaint systems. The authors use data from the New York City 311 system to explore socio-spatial disparities in complaint behavior and how these disparities affect the fairness of data-driven decisions.

[1] Constantine, R., Constable, G., & Gaffney, D. (2016). "An Introduction to Open Science: Towards Transparency and Reproducibility in Geosciences Research". *EOS, Transactions, American Geophysical Union*, 97. https://doi.org/10.1029/2016EO056689

[2] Constantine E. Kontokosta and Bo Yeong Hong, "Bias in smart city governance: How socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions," *Journal of Planning Education and Research* 39, no. 3 (2019): 292-304, https://doi.org/10.1177/0739456X18825076.

The authors find that complaint behavior is not evenly distributed across different neighborhoods and demographic groups, with some areas and groups being more likely to file complaints than others. The authors argue that this can lead to biased data-driven decisions, as areas with lower complaint rates may be underrepresented in the data used for decision-making. The research highlights the importance of inclusive and representative data in smart city governance. The authors suggest that policymakers should address the underlying socio-spatial disparities in complaint behavior by improving access to 311 systems and promoting community engagement. Overall, the paper provides valuable insights into potential bias in smart city governance and the importance of equitable and inclusive data-driven decision-making processes. The authors' findings have important implications for policymakers, urban planners, and researchers working in the field of smart city governance.

# 3. Data

We used multiple open-source datasets throughout our analysis, provided by NYC Open Data Portal or the U.S. Census Bureau and American Community Survey. Please see Table 1 for the data sources' detailed descriptions and primary columns.

In order to produce the best results from our sentiment analysis, we needed sufficient amounts of natural language similar to the length of short sentences. Unfortunately, the three-tiered complaint hierarchy in the 311 service requests dataset does not provide enough description. Thus, we introduced a secondary dataset, 311 Call Center Inquiries, which consists of a similar three-tiered hierarchy that we found to be more descriptive.

To connect our severity metric to specific geographies and communities across the city, we use the 311 service requests dataset because this dataset contains latitude and longitude points for each unique request, whereas 311 call center inquiries do not.

We used the U.S. Census Bureau and ACS data to connect these geographies to socioeconomic metrics such as racial and ethnic population densities, income, poverty rates, and educational attainment. In addition to these socioeconomic predictors, we used NYC Open Data Portal to obtain aggregated crime complaints per capita and the street density for each Neighborhood Tabulation Area. We decided to include these metrics in our analysis because the propensity to call the NYPD and complain about a crime and the propensity to submit a 311 request are intimately related. In addition, street density may influence the propensity to utilize 311 services and the type of complaints, i.e., street and sidewalk-related complaints.

| Data Source | Description | Primary Columns |
|---|---|---|
| **Primary Data Sources** | | |
| 311 Service Requests | Contains all 311 service requests from 2010 to present day with location of request. Requests can be placed via phone, online or application.<br>This data will be used to aggregate complaint severity to neighborhoods across New York City. | ● **agency_name:** responding city agency<br>● **complaint_type:** first level of a hierarchy identifying the topic of the incident or condition<br>● **descriptor:** dependent on the Complaint Type, and are not always required<br>● **location:** geospatial location point of incident |
| 311 Call Center Inquiries | Contains information on all agent-handled calls to the City's 311 information line, including date, time and topic.<br>This data will be used for sentiment analysis and to obtain complaint severity. | ● **agency_name:** responding city agency<br>● **inquiry_name:** first level of a hierarchy identifying the topic the call center representative used to resolve inquiry<br>● **brief_descriptor:** brief description of complain |
| U.S. Census Bureau and ACS | The U.S. Census Bureau conducts a decennial census and this includes data on population, demographics, housing, employment and more<br>ACS includes yearly data on demographics, economic, social and housing characteristics. . | ● **Total Population:** Total population for each census tract<br>● **Below Poverty Line:** Count of population below the poverty line<br>● **Educational Attainment:** Count of population who attained a Bachelor's degree<br>● **Race:** Included population by race - White, Black or African American, Hispanic and more |
| NYPD Complaint Data | This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to present day | ● **cmplnt_num:** unique id for each complaint<br>● **cmplnt_fr_dt:** Date of occurrence<br>● **lat_lon:** geospatial location point of incident |
| Digital City Map Street Center Line | The Digital City Map (DCM) data represents street lines and other features shown on the City Map, which is the official street map of the City of New York. | ● **geometry:** geospatial multilinestring geometry of streets center line |

*Table 1. Data Sources, descriptions and primary column names used throughout our research and analysis*

# 4. Methods

The first step in our research is to create the severity metric using Python's Natural Language Toolkit (NLTK) to process and analyze the text of the unique 311 complaints from 2019. Using the Socrata API querying language, we retrieved the ~ 1980 unique combinations of an agency name, inquiry name, and brief description from the 311 call center inquiries. Since NLTK's pre-trained sentiment analyzer, VADER, is best suited for short sentences, we combined the three-tier complaint hierarchies into one string to serve as our input.

Before inputting the complaint text into VADER, we pre-processed the text to standardize the text across all inputs. We ensured all punctuation was removed, words were lowercase, and all English stop words were removed. Stopwords are common words, like articles, pronouns, prepositions, and conjunctions[3]. Removing stop words is an important step in language processing in order to remove unnecessary and unimportant text articles. Once the complaints' text have been pre-processed, we input them into VADER which outputs four scores, negative, positive, neutral and compound. Negative, positive, and neutral are the percent of text that corresponds to negative, positive and neutral words, respectively. Compound is the normalized sum of the positive, negative, and neutral scores so that it is between -1, most extreme negative, and +1, most extreme positive[4].

The most taxing portion of this analysis was connecting the call center inquiries to the 311 service requests because there is no standard key between the two datasets, and the data structure and content differ. As a result, there

[3] Real Python. (n.d.). Python NLTK Sentiment Analysis – Text Classification. Retrieved May 3, 2023, from https://realpython.com/python-nltk-sentiment-analysis/#using-nltks-pre-trained-sentiment-analyzer

[4] Analytics India Magazine. (2019, January 14). Sentiment Analysis Made Easy Using VADER. Retrieved May 3, 2023, from https://analyticsindiamag.com/sentiment-analysis-made-easy-using-vader/#:~:text=The%20compound%20score%20is%20the,the%20positivity%20of%20the%20text.

was much manual effort on our part to map as many of the unique call center inquiries to the unique complaint types in the 311 service requests dataset as possible. Ultimately, we could map 775 unique call complaints to 106 unique 311 service request complaint types. However, the subset of call complaints that were successfully mapped to 311 service requests seems to follow the same distribution as the entire dataset (see Figure 1.), so we felt confident in using these complaints' compound scores as our severity metric.

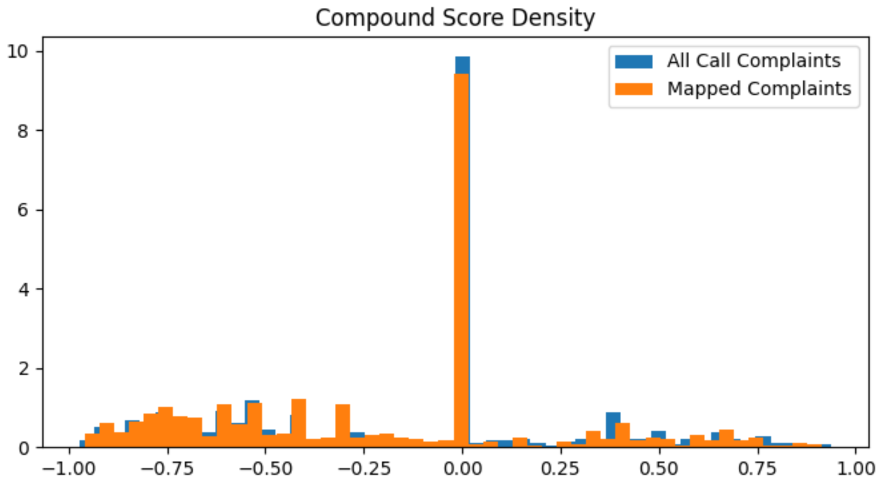

*Figure 1. Density histogram of compound score for all call complaints (blue) and complaints that were successfully mapped to 311 service requests (orange)*

Once we created a dictionary that mapped the unique 311 service request complaint types to the associated compound score, we retrieved all complaints from 2019, mapped them to their associated compound score, and aggregated them to NTAs. As a result, we obtained the following regressors at the NTA geography level: racial and ethnic population density, median income, poverty rate, educational attainment, and road density.

We conducted a multivariate linear regression to pinpoint which regressors were statistically significant in predicting 311 complaint severity. Then, we built a decision tree, optimizing for tree depth and leaf node parameters to interpret which variables are more or less impactful in our prediction.

# 5. Results

Overall, the results obtained from VADER aligned with our intuitions for the most part. Figure 2 shows the top ten negative and positive call complaint categories based on compound score. These figures show that complaints involving abuse, assault, and emergency are the most negative, except for the building code-related categories we assume to be outliers. Additionally, the most positive complaints are all informational requests involving public assistance programs which intuitively would be among the least severe complaint types.

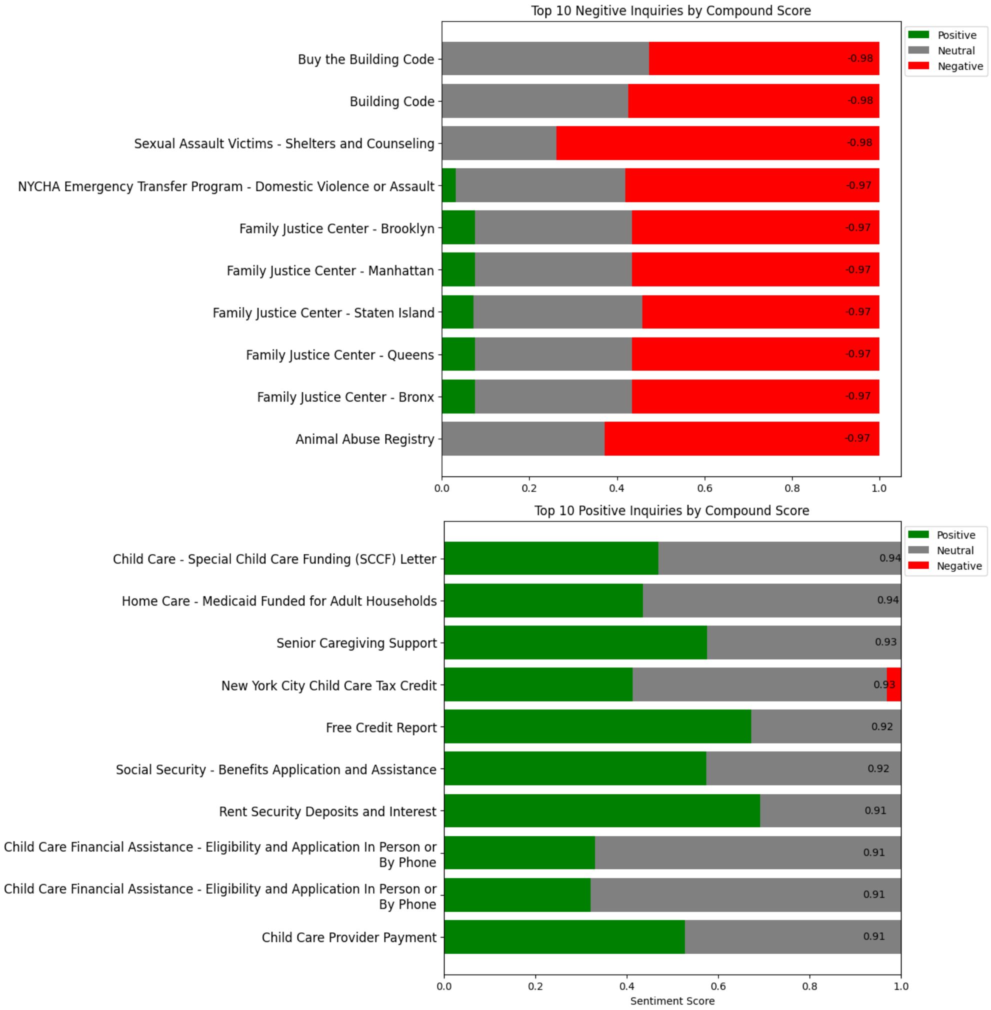

*Figure 2. Top 10 most negative (top) and positive (bottom) 311 call center inquiries by VADER compound score*

After mapping the call complaints to the 311 service requests, we found that elder abuse, school bullying, drug activity, and accident-related complaints were among the most negative. Benefits-related complaints remained the top positive complaint categories. Figure 3 shows the sentiment analysis results spatially for Neighborhood Tabulation Areas (NTAs) across the five boroughs. We calculated the average compound score, weighted on the number of complaints, for each neighborhood (left) and normalized this metric to zero and one (right). Zero corresponds to the most negative complaints, and one is the most positive. The entire borough of Manhattan tends to submit complaints with less negative compound scores. Parts of Brooklyn, mainly DUMBO, Downtown Brooklyn, Cobble Hill, Bedford-Stuyvasent, Williamsburg, and Greenpoint areas, fall within the fourth and fifth percentile of compound scores. The outer neighborhoods of Brooklyn, Bronx, Queens, and Staten Island are within the top 50% for the most severe or negative compound score.

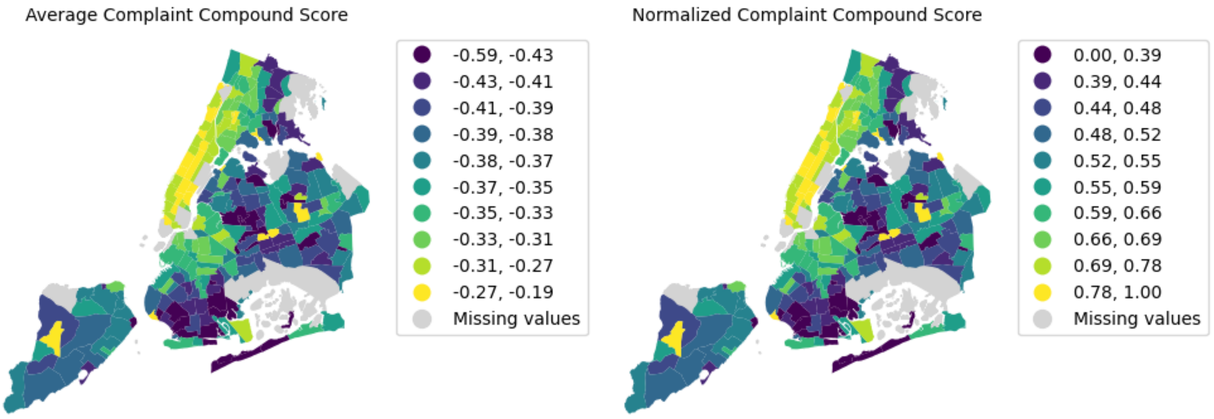

*Figure 3. Weighted average compound score (left) and normalized (right) for NTAs across New York City.*

These trends are consistent when compared to the various regressors we are interested in, especially regarding median income. Figure 4 shows the spatial distribution of four primary variables: median income, poverty rate, white and non-Hispanic population density, and New York Police Department crime complaints per capita. The areas with the lowest complaint severity tend to have the highest median income.

In order to pinpoint which of our many regressors were statistically significant in measuring 311 complaint severity, we initially ran a multivariate linear regression. The results showed that only four out of the original thirteen regressors are significant (Table 2).

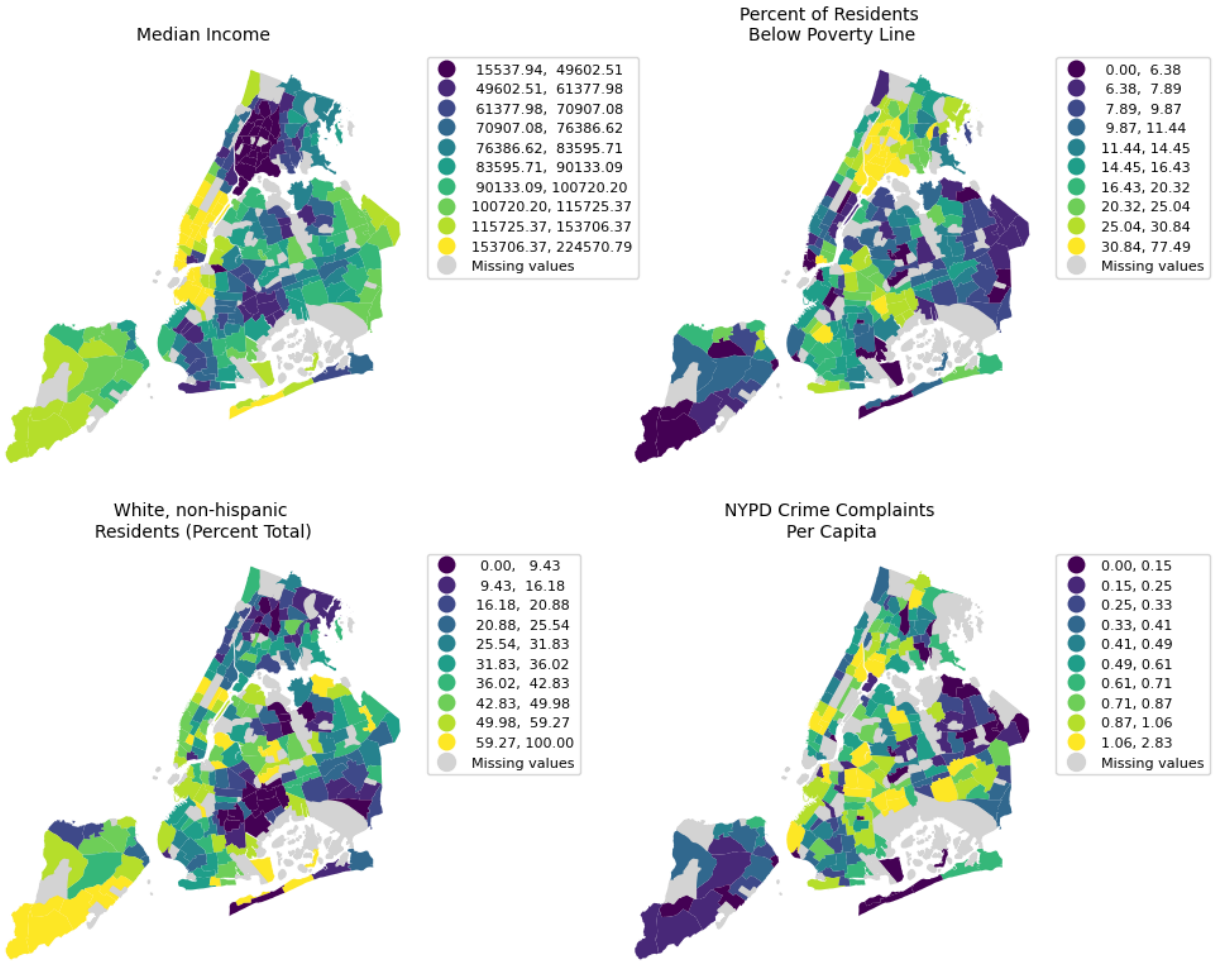

*Figure 4. Spatial distribution of main regressors/predictors for 311 complaint severity*

| Variable/Regressor | P-Value | Variable/Regressor | P-Value |
|---|---|---|---|
| Median Income | 0.000 | Poverty Rate | 0.000 |
| Educational Attainment (Bach) | 0.002 | Racial and Ethnic Population Density* | [0.208, 0.794] |
| Street Density | 0.637 | NYPD Complaints | 0.017 |

*Table 2. Multivariate linear regressor variables and p-values. Red indicates insignificant and green indicates significant.*

It is important to note that all racial and ethnic population densities were insignificant, except Hispanic or Latino and two or more races. These results were somewhat consistent with the results from our decision tree feature importance results. Figure 5 shows that the most important feature in our model is median income, with poverty rate and educational attainment second and third respectively. However, feature importance output shows that street density is also important in predicting 311 complaint severity, which is contradictory to the multivariate linear regression results seen in Table 2. It is important to note that the top three most important regressors are intimately correlated with one another, however, since decision trees are not affected by multicollinearity and covariance[5] we will input all regressors into our model.

[5] Mane, P. (2019, March 21). Multicollinearity in Tree-Based Models. Medium. Retrieved May 3, 2023, from https://medium.com/@manepriyanka48/multicollinearity-in-tree-based-models-b971292db140

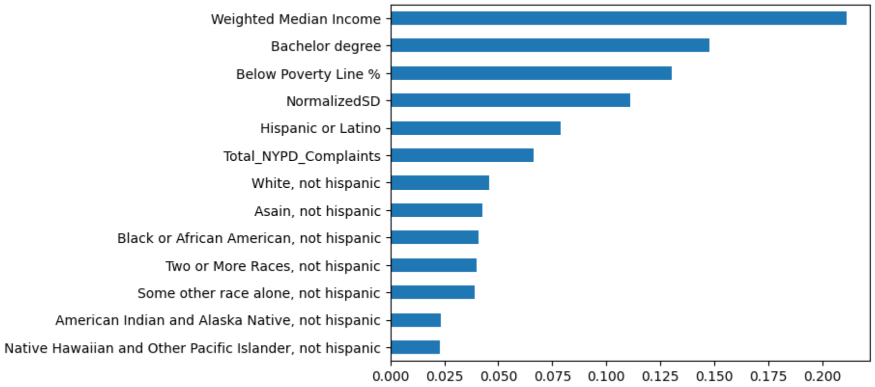

*Figure 5. Feature importance of regressors in decision tree model shows street density (NormalizedSD) as the fourth most important feature, ranked over NYPD complaints per capita.*

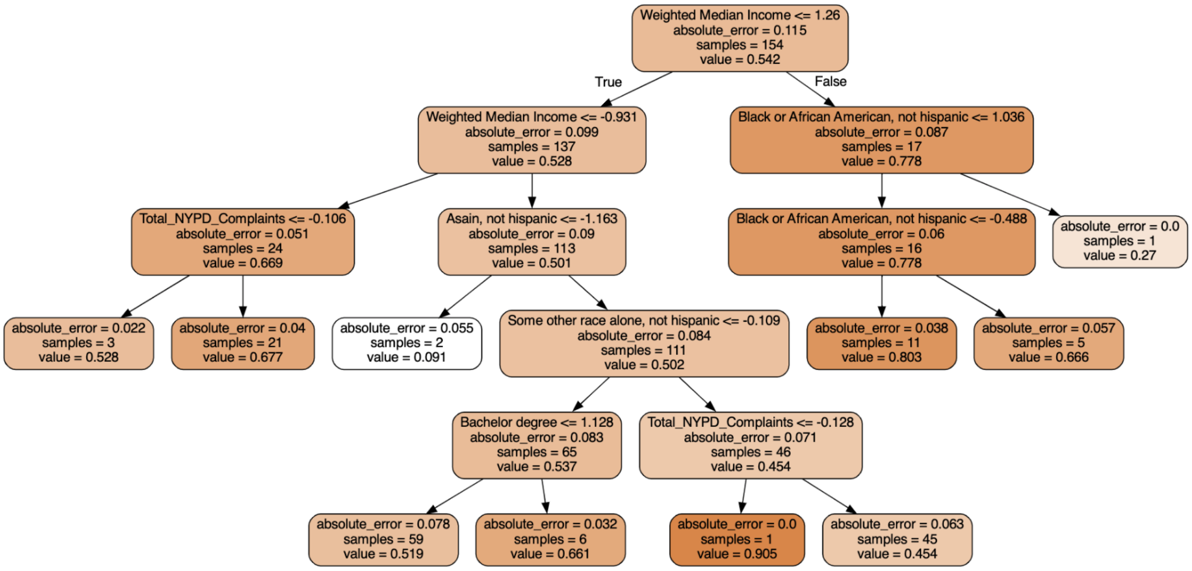

*Figure 6. Decision tree produced after hyperparameter tuning. Optimal tree was built with {'criterion': 'absolute_error', 'max_depth': 5, 'max_leaf_nodes': 10, 'min_samples_leaf': 1, 'min_samples_split': 5}.*

Without hyperparameter tuning, our decision tree regressor reported an in-sample r-square value of 1 and sample r-square of 0.09+/-0.05, a suboptimal result. To increase the out-of-sample accuracy, we tuned our parameters using GridSearchCV. The optimal parameters used to initialize the Decision Tree Regressor are:

- Criterion: In this case, 'absolute_error' is used, which means that the absolute error metric is used to evaluate the splits. This means that the splits will be chosen to minimize the absolute difference between the predicted and actual values of the target variable.
- Max_depth: A deeper tree can capture more complex relationships in the data but may also lead to overfitting. Here, the maximum depth is set to 5, meaning the tree can have a maximum of five levels.
- Min_samples_split: This parameter can be used to control the complexity of the model and prevent overfitting. In this case, the minimum number of samples required to split a node is set to 5.
- Max_leaf_node: It controls the maximum number of leaf nodes that the decision tree can have. The default value is ‘None’ but here to get the most optimal tree, we have set it to 10 as it strikes a balance between interpretability and predictive accuracy.
- Min_sample_leaf: This hyperparameter controls the minimum number of samples required to form a leaf node. Here, it is set as 1 in order to have a more granular decision tree that captures more details in the data.

The tree was trained using a dataset of 154 samples over 100 random 80-20 train-test splits and reported an in-sample r-square of 0.65 +/- 0.03 and an out-of-sample r-square of -0.08 +/- 0.32. While the average out-of-sample r-square over the 100 random samples is relatively poor, the standard deviation is significant, putting our results in a more reasonable range of accuracy. The resulting decision tree can be found in Figure 6. Each split node in the tree shows the target variable's predicted value, which is our severity score. The absolute error of the prediction is also displayed, representing the average absolute difference between the predicted value and the actual value for the samples in that split. Feature and threshold used to split the data, the number of samples that fall into that split, the predicted probability of the positive class, and the absolute error of the prediction

The tree has a root node that splits the dataset based on the "Weighted Median Income" feature. It is important to note that we standardized the regressors using sklearn's StandardScalar method before building out the decision tree, so the values in the tree do not correspond with the raw metrics seen previously in the report. For example, based on the first split, if the neighborhood has less than or equal to 1.26 in standardized median income, the sample goes down the left branch, and if it is more significant than 1.26, the sample goes down the right branch. This process recursively until each leaf node contains a subset of the data with a homogeneous class distribution.

There are two main trends and relationships seen in the decision tree in Figure 6. The first is that high-income neighborhoods with less than average population density of Black or African American residents have low 311 complaint severity. The second is that low-income neighborhoods, with less than average NYPD complaints per capita have higher 311 complaint severity. These two trends have been previously stipulated by researchers and urban scientists, particularly in regards to racial and income indicators[1].

We investigated ensemble methods as a means of increasing accuracy in our regression model. However, after exploring random forests, we found that it did not provide any further insight into the relationship between our regressors and predictor. Ultimately, we decided to stick with our decision tree model and focus on fine-tuning its hyperparameters.

# 6. Conclusion

In conclusion, we explored using sentiment analysis and decision trees for measuring and predicting 311 service request severity. We began with a discussion of the importance of accurate severity prediction in optimizing the allocation of city resources and the limitations of current methods. Then, we introduced the concept of sentiment analysis, which was used to extract sentiment scores from the text description of service requests. These scores were then used with other features, such as racial and ethnic population density, poverty rate, and educational attainment to train a decision tree model to predict service request severity.

Our results showed that the sentiment score extracted from the text descriptions of service requests was intuitive and is a valuable indicator of the level of urgency and seriousness of the issue. However, we recommend more research be done about the use of sentiment analysis as a method of measuring severity. We believe there is an opportunity for local agencies to invest in recording 311 call complaints to use with a sentiment analyzer trained specifically on complaints, not social media.

The decision tree model, which predicted sentiment scores across neighborhoods given various socioeconomic variables, had a suboptimal performance. However, the decision tree was extremely helpful in interpreting the relationship our variables have with determining complaint severity. We believe the performance of this model would increase given smaller geographies, i.e. more samples, and we trained our own sentiment analyzer specific to 311 complaint natural language data.

Overall,  our results suggest that sentiment analysis and decision trees are promising tools for predicting 311 service request severity and have the potential to improve the efficiency and effectiveness of city services significantly.

# 7. Appendix

**Github Repository -** https://github.com/emj1020/311-sentiment-analysis